OPEN ACCESS Freely available online

PLOS ONE

Additive and Subtractive Scrambling in Optional Randomized Response Modeling

Zawar Hussain[1]*, Mashail M. Al-Sobhi[2]*, Bander Al-Zahrani[3]*

1 Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan, **2** Department of Mathematics, Umm Alqura University, Makkah, Saudi Arabia, **3** Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract

This article considers unbiased estimation of mean, variance and sensitivity level of a sensitive variable via scrambled response modeling. In particular, we focus on estimation of the mean. The idea of using additive and subtractive scrambling has been suggested under a recent scrambled response model. Whether it is estimation of mean, variance or sensitivity level, the proposed scheme of estimation is shown relatively more efficient than that recent model. As far as the estimation of mean is concerned, the proposed estimators perform relatively better than the estimators based on recent additive scrambling models. Relative efficiency comparisons are also made in order to highlight the performance of proposed estimators under suggested scrambling technique.

Citation: Hussain Z, Al-Sobhi MM, Al-Zahrani B (2014) Additive and Subtractive Scrambling in Optional Randomized Response Modeling. PLoS ONE 9(1): e83557. doi:10.1371/journal.pone.0083557

Editor: Yinglin Xia, University of Rochester, United States of America

Received July 26, 2013; **Accepted** November 5, 2013; **Published** January 8, 2014

Funding: The work has been funded by the Institute of Scientific Research and Revival of Islamic Heritage at Umm Al-Qura University (grant # 43305030). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: zhlangah@yahoo.com (ZH); mmsobhi@uqu.edu.sa (MA); bmalzahrani@kau.edu.sa (BA)

Introduction

To procure reliable data on stigmatizing characteristics, Warner [1] introduced the notion of randomized response technique where the respondent himself selects randomly one of the two complementary questions on probability basis. Greenberg et al. [2] extended the Warner's [1] work to collect the data on quantitative stigmatizing variables. Since then, several authors have worked on quantitative randomized response models including, Eichhorn and Hayre [3], Gupta and Shabbir [4], Gupta and Shabbir [5], Bar-Lev et al. [6], Gupta et al. [7], Hussain and Shabbir [8], Saha [9], Chaudhuri [10], Hussain and Shabbir [11] and references therein. Quantitative randomized response models are classified into fully (Eichhorn and Hayre [3]), partial (Gupta and Shabbir [5]), Bar-Lev et al. [6]) and optional randomized response models (Gupta et al. [4]), Gupta et al. [7], Huang [12]). In a fully randomized response models all the responses are obtained as scrambled responses. In a partial randomized response model a known proportion of respondents is asked to report their actual responses while the others report scrambled responses.

Our focus in this article is on ORRMs only. The notion of ORRM started with Gupta et al. [4]. The concept of ORRM is based on the respondent's perception about sensitivity of the variable of interest. Using ORRM, a respondent can report the truth (or scramble his/her response) if he/she perceives the study variable as non sensitive (sensitive) to him/her. The proportion of respondents reporting the scrambled response is unknown, and is termed as the sensitivity level of the study variable. Gupta et al. [4] used multiplicative ORRM and provided unbiased (biased) estimator of mean (sensitivity). Moreover, Gupta et al. [4] ORRM requires approximation in order to derive the variances of the estimators. In Gupta et al. [4] ORRM, simultaneous estimation of mean and sensitivity is not possible. To avoid approximation, Gupta et al. [7], Huang [12], Gupta et al. [13] and Mehta et al. [14] proposed ORRMs to provide unbiased estimators of mean and sensitivity level. Gupta et al. [7] and Huang [12] are the one-stage ORRMs, Gupta et al. [13] is a two-stage ORRM whereas Mehta et al. [14] is a three-stage ORRM. Gupta et al. [7], Gupta et al. [13] and Mehta et al. [14] used additive scrambling whereas Huang [12] used a linear combination of additive and multiplicative scrambling. Further, Gupta et al. [15] observed that additive scrambling yields more precise estimators than a linear combination of additive and multiplicative scrambling by Huang [12]. Also, Gupta et al. [16] observed that in Gupta et al. [13] two-stage ORRM a large value of truth parameter (T) is required when the study variable is highly sensitive. Motivated by the advocacy of additive scrambling and requirement of larger value of truth parameter (T), Mehta et al. [14] proposed a three stage ORRM by introducing a forced scrambling parameter (F). Mehta et al. [14] established the better performance of estimator of mean but did not discuss the performance of sensitivity estimator. As far as the estimation of mean is concerned, Mehta et al. [14] ORRM can be further improved by using a multi-stage randomization but it results in a poor estimation of sensitivity level.

All of the ORRMs mentioned above share a common feature of splitting the total sample into two subsamples. We base our proposals on two strategies: (i) taking two subsamples and making use of additive scrambling in one subsample and subtractive scrambling in the other, and (ii) drawing a single sample and collecting two responses from each respondent through additive and subtractive scrambling. Through our strategies, we plan to improve Mehta et al. [14] ORRM for estimating the mean. As far as estimation of mean is concerned, we show that the proposed ORRM is better than Mehta et al. [14], Huang [12] and Gupta et

al. [13] ORRMs. We show that there is no need of large value of the parameter (T or F) when the study variable is either low, moderately or highly sensitive. In addition, we also propose an estimator of the variance of the study variable.

We now briefly discuss three of the background ORRMs, namely, the Mehta et al. [14], Huang [12] and Gupta et al. [13].

Mehta et al. [14] ORRM

Assume that the interest lies in unbiased estimation of the mean μ_X and the sensitivity level W of the study variable X. Let $D_i,(i=1,2)$ be the unrelated scrambling variable. Two independent subsamples of size $n_i(i=1,2)$, are drawn from the population through simple random sampling with replacement such that $n_1+n_2=n$, the total sample size required. In i^{th} subsample, a fixed predetermined proportion (T) of respondents is instructed to tell the truth and a fixed predetermined proportion (F) of respondents is instructed to scramble additively their response as $(X+D_i)$. The remaining proportion $(1-T-F)$ of respondents have an option to scramble their response additively if they consider the study variable sensitive. Otherwise, they can report the true response X. Let $\mu_{D_i}=\theta_i$, be the known mean, and $\sigma^2_{D_i}=\delta_i^2$, be the known variance of the positive-valued random variable $D_i(i=1,2)$. The optional randomized response from j^{th} respondent in the i^{th} subsample is given by:

$$Z_{ij}=\alpha_j X_j+\beta_j\left(X_j+D_{ij}\right)+\left(1-\alpha_j-\beta_j\right)\left\{\left(1-\ Y_j\right)X_j+Y_j\left(X_j+D_{ij}\right)\right\}, \tag{1}$$

where $i=1,2.,j=1,2,...,n_i,\ Y_j\sim Bernoulli(W),\ \alpha_j\sim Bernoulli(T)$ and $\beta_j\sim Bernoulli(F)$. The expectation of the sample response Z_{ij} from i^{th} sample is given by:

$$E\left(Z_{ij}\right)=\mu_X+(F+(1-T-F)W)\theta_i.$$

Taking $\bar{Z}_1$ and $\bar{Z}_2$ as the observed means from the two subsamples, Mehta et al. [14] proposed the following estimators of μ_X and W.

$$\hat{\mu}_{XM}=\frac{\theta_1\bar{Z}_2-\theta_2\bar{Z}_1}{(\theta_1-\theta_2)},\ \theta_1\neq\theta_2 \tag{2}$$

$$\hat{W}_M=\frac{1}{(1-T-F)}\left(\frac{\bar{Z}_2-\bar{Z}_1}{(\theta_2-\theta_1)}-F\right),\ T+F\neq1,\theta_1\neq\theta_2. \tag{3}$$

The variances of estimators in (2) and (3) are given by:

$$Var(\hat{\mu}_{XM})=\frac{1}{(\theta_1-\theta_2)^2}\left(\theta_2^2\frac{\sigma^2_{Z_1}}{n_1}+\theta_1^2\frac{\sigma^2_{Z_2}}{n_2}\right) \tag{4}$$

$$Var\left(\hat{W}_M\right)=\frac{1}{(1-T-F)^2(\theta_1-\theta_2)^2}\left(\frac{\sigma^2_{Z_1}}{n_1}+\frac{\sigma^2_{Z_2}}{n_2}\right), \tag{5}$$

where

$$\sigma^2_{Z_i}=\sigma^2_X+(F+(1-T-F)W)\{1-(F+(1-T-F)W)\}\theta_i^2+(F+(1-T-F)W)\delta_i^2,\quad (i=1,2). \tag{6}$$

Gupta et al. [13] ORRM

It is interesting to note that for $F=0$, the Mehta et al. [14] ORRM reduces to Gupta et al. [13] ORRM. . Let Z'_{ij} be the optional scrambled response from j^{th} respondent in the i^{th} subsample then taking $F=0$ in (1)–(5), unbiased estimators and their variances are given by:

$$\hat{\mu}_{XG}=\frac{\theta_1\bar{Z}'_2-\theta_2\bar{Z}'_1}{(\theta_1-\theta_2)},\ \theta_1\neq\theta_2 \tag{7}$$

$$\hat{W}_G=\frac{\bar{Z}'_1-\bar{Z}'_2}{(1-T)(\theta_1-\theta_2)},\ \theta_1\neq\theta_2,T\neq1. \tag{8}$$

$$Var(\hat{\mu}_{XG})=\frac{1}{(\theta_1-\theta_2)^2}\left(\theta_2^2\frac{\sigma^2_{Z'1}}{n_1}+\theta_1^2\frac{\sigma^2_{Z'2}}{n_2}\right) \tag{9}$$

$$Var\left(\hat{W}_G\right)=\frac{1}{(1-T)^2(\theta_1-\theta_2)^2}\left(\frac{\sigma^2_{Z'1}}{n_1}+\frac{\sigma^2_{Z'2}}{n_2}\right), \tag{10}$$

where

$$\sigma^2_{Z'_i}=\sigma^2_X+W(1-T)\{1-W(1-T)\}\theta_i^2+W(1-T)\delta_i^2.$$

Huang [12] ORRM

Each respondent in the i^{th} subsample is provided with two randomization devices which generate two independent random variables, say S_i and D_i, from some pre-assigned distributions. The respondent chooses randomly by himself one of the following two options: (a) report the true response X (if you do not feel the study variable sensitive), or (b) report the scrambled response S_iX+D_i (if you feel the study variable sensitive). Let $\mu_{S_i}=1$, be the known mean, and $\sigma^2_{S_i}=\gamma_i^2$, be the known variance of the positive-valued random variables S_i. The optional randomized response Z''_{ij} from j^{th} respondent in the i^{th} subsample is given by:

$$Z''_{ij}=\left(1-\ Y_j\right)X_j+Y_j\left(S_{ij}X_j+D_{ij}\right), \tag{11}$$

The expectation of sample response Z''_{ij} from i^{th} sample is given by:

$$E\left(Z''_{ij}\right)=(1-W)\mu_X+W\{\mu_X+\theta_i\}=\mu_X+W\theta_i,$$

since $\mu_{S_i}=1$. Huang [2] proposed the following estimators of μ_X and W.

$$\hat{\mu}_{XH}=\frac{\theta_1\bar{Z}_2''-\theta_2\bar{Z}_1''}{(\theta_1-\theta_2)},\ \theta_1\neq\theta_2 \tag{12}$$

$$\hat{W}_H=\frac{\bar{Z}_1''-\bar{Z}_2''}{(\theta_1-\theta_2)},\ \theta_1\neq\theta_2, \tag{13}$$

where $\bar{Z}_1''$ and $\bar{Z}_2''$ are the observed means from the two subsamples. The variances of estimators in (12) and (13) are given by:

$$Var(\hat{\mu}_{XH})=\frac{1}{(\theta_1-\theta_2)^2}\left(\theta_2^2\frac{\sigma^2_{Z_1''}}{n_1}+\theta_1^2\frac{\sigma^2_{Z_2''}}{n_2}\right) \tag{14}$$

$$Var(\hat{W}_H)=\frac{1}{(\theta_1-\theta_2)^2}\left(\frac{\sigma^2_{Z_1''}}{n_1}+\frac{\sigma^2_{Z_2''}}{n_2}\right), \tag{15}$$

where

$$\sigma^2_{Z''_i}=\sigma^2_X+W(\mu^2_X+\sigma^2_X)\gamma_i^2+W(1-W)\theta_i^2+W\delta_i^2.$$

Proposed Procedures

In this section, we propose split sample and double response approaches using Mehta et al. [14] ORRM.

Split sample approach

Unlike Mehta et al. [14], in the proposed procedure, we use an additive scrambling in one subsample and subtractive scrambling in the other. All the other procedure is same as that of Mehta et al. [14]. Let R_{1j} and R_{2j} be response from $j^{th}(j=1,2,...,n_i)$ respondent selected in the $i^{th}(i=1,2)$ sample, then R_{1j} and R_{2j} can be written as:

$$\begin{aligned}R_{1j}=&\alpha_jX_j+\beta_j(X_j+D_{1j})+(1-\alpha_j-\beta_j)\\&\{(1-Y_j)X_j+Y_j(X_j+D_{1j})\}.\end{aligned} \tag{16}$$

$$\begin{aligned}R_{2j}=&\alpha_{1j}X_j+\beta_j(X_j-D_{2j})+(1-\alpha_j-\beta_j)\\&\{(1-Y_j)X_j+Y_j(X_j-D_{2j})\}.\end{aligned} \tag{17}$$

The expected responses from the two subsamples are given by:

$$E(R_{1j})=\mu_X+(F+(1-T-F)W)\theta_1. \tag{18}$$

$$E(R_{2j})=\mu_X-(F+(1-T-F)W)\theta_2. \tag{19}$$

Solving (18) and (19), we get:

$$\mu_{XZ}=\frac{\theta_1E(R_{2j})+\theta_2E(R_{1j})}{(\theta_1+\theta_2)}.$$

$$W_Z=\frac{1}{(1-T-F)}\left(\frac{E(R_{1j})-E(R_{2j})}{(\theta_1+\theta_2)}-F\right).$$

Estimating $E(R_{1j})$ and $E(R_{2j})$ by the respective sample means $\bar{R}_1$ and $\bar{R}_2$, unbiased estimators of μ_X and W are proposed as:

$$\hat{\mu}_{XZ}=\frac{\theta_1\bar{R}_2+\theta_2\bar{R}_1}{(\theta_1+\theta_2)}. \tag{20}$$

$$\hat{W}_Z=\frac{1}{(1-T-F)}\left(\frac{\bar{R}_1-\bar{R}_2}{(\theta_1+\theta_2)}-F\right). \tag{21}$$

Unbiasedness of $\hat{\mu}_{XZ}$ and $\hat{W}_Z$ can be easily established through (18) and (19). The variances of $\hat{\mu}_{XZ}$ and $\hat{W}_Z$ are given by :

$$Var(\hat{\mu}_{XZ})=\frac{1}{(\theta_1+\theta_2)^2}\left(\theta_2^2\frac{\sigma^2_{R_1}}{n_1}+\theta_1^2\frac{\sigma^2_{R_2}}{n_2}\right), \tag{22}$$

$$Var(\hat{W}_Z)=\frac{1}{(1-T-F)^2(\theta_1+\theta_2)^2}\left(\frac{\sigma^2_{R_1}}{n_1}+\frac{\sigma^2_{R_2}}{n_2}\right), \tag{23}$$

where $\sigma^2_{R_i}=\sigma^2_{Z_i}$.

It is important to note that subtractive scrambling in the second subsample is same as the additive scrambling if $-D_2$ is viewed as the new scrambling variable. We anticipate two advantages by calling it subtractive scrambling. Firstly, it is easier just to subtract a constant (randomly chosen by the respondent) from the actual response on sensitive variable. Second advantage is a psychological one in nature. Perhaps, due to social desirability, a typical respondent would like to report smaller response in magnitude. In other words, respondents would be happy in underreporting, in general. Thus, subtracting a positive constant from the actual response would help satisfying the social desirability of underreporting. Of course, these two advantages are gained in the second subsample only since D_1 and D_2 are positive valued random variables. On average, affect of additive scrambling in one subsample is offset by subtractive scrambling in the other. As a result, parameters are estimated with increased precision.

Theorem 2.1: For $T+F<1$, $\hat{\mu}_{XZ}\sim N(\mu_X,Var(\hat{\mu}_{XZ}))$ and $\hat{W}_Z\sim N(W,Var(\hat{W}_Z))$.

Proof: Since $\hat{\mu}_{XZ}$ and $\hat{W}_Z$ are the linear combinations of sample means, application of central limit theorem gives the required result.

In view of the fact that $s^2_{R_i}=(n_i-1)^{-1}\sum_{j=1}^{n_i}(R_{ji}-\bar{R}_i)^2$ is an unbiased estimator of $\sigma^2_{R_i}$ we have the following theorems.

Theorem 2.2: An unbiased estimator of $Var(\hat{\mu}_{XZ})$ is given by:

$$\hat{V}ar(\hat{\mu}_{XZ})=\frac{1}{(\theta_1+\theta_2)^2}\left(\theta_2^2\frac{s^2_{R_1}}{n_1}+\theta_1^2\frac{s^2_{R_2}}{n_2}\right).$$

Theorem 2.3: An unbiased estimator of the $Var(\hat{W}_Z)$ is given by:

$$\hat{V}ar(\hat{W}_Z)=\frac{1}{(\theta_1-\theta_2)^2}\left(\frac{s^2_{R_1}}{n_1}+\frac{s^2_{R_2}}{n_2}\right).$$

Proofs: The proofs of the above Theorems (2.2 and 2.3) can easily be provided by utilizing the fact that $E\left(s^2_{R_i}\right)=\sigma^2_{R_i}$.

Theorem 2.4: An unbiased estimator of $Var(Y)$ is given by:

$$\hat{V}ar(Y)=\hat{W}_Z(1-\hat{W}_Z)+\hat{V}ar(\hat{W}_Z).$$

Proof: Applying the expectation operator at $\hat{V}ar(Y)$, we get:

$$E\{\hat{V}ar(Y)\}=E(\hat{W}_Z)-E(\hat{W}^2_Z)+E\{\hat{V}ar(\hat{W}_Z)\}.$$

Then, applying Theorem 2.3, we get:

$$E\{\hat{V}ar(Y)\}=E(\hat{W}_Z)-E(\hat{W}^2_Z)+Var(\hat{W}_Z)$$

$$E\{\hat{V}ar(Y)\}=E(\hat{W}_Z)-\left[Var(\hat{W}_Z)+\{E(\hat{W}_Z)\}^2\right]+Var(\hat{W}_Z)$$

$$E\{\hat{V}ar(Y)\}=W-W^2=W(1-W).$$

Now, we consider the estimation of variance σ^2_X of the sensitive variable X. Provided that $\delta^2_2-\delta^2_1\neq 0$, from (6) we can, after a simple algebra, write that

$$\sigma^2_X=\frac{\delta^2_2\sigma^2_{R_1}-\delta^2_1\sigma^2_{R_2}-(\theta^2_1\delta^2_2-\theta^2_2\delta^2_1)A(1-A)}{(\delta^2_2-\delta^2_1)},$$
$$A=\{F+(1-T-F)W_Z\}.$$

We define unbiased estimators of σ^2_X in the following theorems.

Theorem 2.5: In case when $\delta^2_2-\delta^2_1\neq 0$, an unbiased estimator of σ^2_X is given by:

$$\begin{aligned}\hat{\sigma}^2_{XZ}=\delta^2_2 s^2_{R_1}-\delta^2_1 s^2_{R_2}-(\delta^2_2\theta^2_1-\delta^2_1\theta^2_2)\{F(1-F)+\\ (1-2F)(1-T-F)\hat{W}-(1-T-F)^2\\ (\hat{W}-Var(\hat{W}))\}/(\delta^2_2-\delta^2_1).\end{aligned}\quad(24)$$

Theorem 2.6: In case when $\delta^2_2-\delta^2_1=0$, an unbiased estimator of σ^2_X is given by:

$$\hat{\sigma}^2_{XZ}=\beta\hat{\sigma}^2_{X1}+(1-\beta)\hat{\sigma}^2_{X2},\quad(25)$$

where β is known constant belonging to the interval [0,1], $\hat{\sigma}^2_{Xi}=s^2_{X_i}+\delta^2_i\left(\hat{A}\right)+\hat{A}\left(1-\hat{A}\right)\theta^2_i$ and $\hat{A}=\left\{F+(1-T-F)\hat{W}_Z\right\}$.

Proofs: The above Theorems 2.5 and 2.6 can be proved by noting that $\hat{W}_Z$ and $\left\{\hat{W}^2_Z-\hat{V}ar(\hat{W}_Z)\right\}$ are unbiased estimators of W and W^2 respectively. Taking expectation of (24) and (25), we get $E(\hat{\sigma}^2_{XZ})=\sigma^2_X$.

Double response approach

Without incurring any additional sampling cost, Mehta et al. [14] ORRM may also be improved by taking two responses from each respondent. We take scrambling variables the same as defined in Mehta et al. [14] ORRM. To report the first (second) response, respondents are requested to use additive (subtractive) scrambling with the variable $D_1(D_2)$. Let R'_{1j} and R'_{2j} be the two responses of j^{th} respondent then the two responses can be written as

$$\begin{aligned}R'_{1j}=\alpha_j X_j+\beta_j(X_j+D_{1j})+(1-\alpha_j-\beta_j)\\ \{(1-Y_j)X_j+Y_j(X_j+D_{1j})\}\end{aligned}$$

$$\begin{aligned}R'_{2j}=\alpha_j X_j+\beta_j(X_j-D_{2j})+(1-\alpha_j-\beta_j)\\ \{(1-Y_j)X_j+Y_j(X_j-D_{2j})\}.\end{aligned}$$

It is obvious from (26) and (27) that the true value of sensitive variable X_j cannot be worked out for the respondents feeling study variable sensitive enough. The reported responses of a particular respondent would be same if he/she feels study variable insensitive. In this case, he/she reports true value of study variable both the times. This is not challenging since the respondents feeling study variable insensitive would be willing to dispose their true value on sensitive variable. Thus, it may be concluded that privacy of respondents, feeling study variable sensitive, remains intact. As correctly pointed out by one of the referees, there is extra burden on the respondent if he/she has to report twice. This issue may be tackled by explaining whole the procedures to the respondent before actually obtaining data. He/she must be assured that his/her actual response on sensitive variable cannot be traced back to his/her actual response. Further he/she must be made clear that interest of the study lies in the estimation of parameters only. Moreover, we do not need any additional sampling cost to obtain two responses. Thus, obtaining two responses from a respondent should not be an issue in a particular study.

The expected responses from the j^{th} respondent are same as given by (18) and (19). Thus $E\left(R'_{1j}\right)=E(R_{1j})$ and $E\left(R'_{2j}\right)=E(R_{2j})$. This implies that unbiased estimators of μ_X and W may be suggested as:

$$\hat{\mu}'_{XZ}=\frac{\theta_1\bar{R}'_2+\theta_2\bar{R}'_1}{(\theta_1+\theta_2)}.\quad(28)$$

$$\hat{W}'_Z=\frac{1}{(1-T-F)}\left(\frac{\bar{R}'_1-\bar{R}'_2}{(\theta_1+\theta_2)}-F\right).\quad(29)$$

The variances of $\hat{\mu}'_{XZ}$ and $\hat{W}'_Z$ are given by :

$$Var\left(\hat{\mu}'_{XZ}\right)=\frac{1}{(\theta_1+\theta_2)^2}\left(\theta^2_2\frac{\sigma^2_{R'_1}}{n}+\theta^2_1\frac{\sigma^2_{R'_2}}{n}+\frac{2Cov\left(R'_1,R'_2\right)}{n}\right),\quad(30)$$

$$Var\left(\hat{W}_Z'\right)=\frac{1}{(1-T-F)^2(\theta_1+\theta_2)^2}\left(\frac{\sigma^2_{R_1'}}{n}+\frac{\sigma^2_{R_2'}}{n}-\frac{2Cov\left(R_1',R_2'\right)}{n}\right),$$

where

$$Cov\left(R_1',R_2'\right)=E\left(R_1'R_2'\right)-E\left(R_1'\right)E\left(R_2'\right)$$

$$Cov\left(R_1',R_2'\right)=\sigma^2_X-\{F+(1-T-F)\}[1-\{F+(1-T-F)\}]\theta_1\theta_2.$$

In some studies, interest of researchers lies in estimating $\mu_X$rather than the sensitivity level W of variable X while it is of major interest in other studies. Following Huang [12], we define a linear combination of $Var(\hat{\mu}_{XZ})$ and $Var\left(\hat{W}_Z\right)$ in order to find the optimum allocation of sample size. Thus, depending upon the interest of researchers, optimum subsample sizes can be obtained.

Consider,

$$Var\left(\hat{\mu}_{XZ},\hat{W}_Z\right)=kVar(\hat{\mu}_{XZ})+(1-k)Var\left(\hat{W}_Z\right),\qquad k\in[0,1].$$

Using Lagrange approach to minimize $Var\left(\hat{\mu}_{XZ},\hat{W}_Z\right)$ under the restriction that $\sum_{i=1}^{2}n_i=n$, we get:

$$n_1=n\frac{\sqrt{\sigma^2_{R_1}\left\{k\left(\theta_2^2-1\right)+1\right\}}}{\sqrt{\sigma^2_{R_1}\left\{k\left(\theta_2^2-1\right)+1\right\}}+\sqrt{\sigma^2_{R_2}\left\{k\left(\theta_1^2-1\right)+1\right\}}},$$

and

$$n_2=n\frac{\sqrt{\sigma^2_{R_2}\left\{k\left(\theta_1^2-1\right)+1\right\}}}{\sqrt{\sigma^2_{R_1}\left\{k\left(\theta_2^2-1\right)+1\right\}}+\sqrt{\sigma^2_{R_2}\left\{k\left(\theta_1^2-1\right)+1\right\}}}.$$

With these optimum sample sizes, the minimum value of $Var\left(\hat{\mu}_{XZ},\hat{W}_Z\right)$ is given by:

$$Min.Var\left(\hat{\mu}_{XZ},\hat{W}_Z\right)=\frac{\left[\sqrt{\sigma^2_{R_1}\left\{k\left(\theta_2^2-1\right)+1\right\}}+\sqrt{\sigma^2_{R_2}\left\{k\left(\theta_1^2-1\right)+1\right\}}\right]^2}{n(\theta_1+\theta_2)^2}.$$

In practice, $\sigma^2_{R_i}$ is unknown and the optimum allocation of sample sizes cannot be made. Following Murthy [17], the unknown values of $\sigma^2_{R_i}$ can be estimated from pilot surveys, past experience or simply an intelligent guess can be made about $\sigma^2_{R_i}$.

Privacy Protection Discussion

There are many privacy measures suggested by different authors. We take $E(Z_i-X_i)^2$ as the measure of privacy. This measure of privacy is proposed by Zaizai et al. [18]. A given model is taken as more protective against privacy if $E(Z_i-X_i)^2$ is higher. For a model providing privacy protection to some extent $E(Z_i-X_i)^2>0$. On the other hand, if a model does not provide any privacy $E(Z_i-X_i)^2=0$. For a given model, the larger the $E(Z_i-X_i)^2$, the larger the privacy provided by the model.

The measures of privacy for Mehta et al. [14] ORRM are given by $W(1-T-F)\left(\theta_1^2+\delta_1^2\right)$ and $W(1-T-F)\left(\theta_2^2+\delta_2^2\right)$ in the first and second subsamples, respectively. Similarly for Gupta et al. [13] model it is $W(1-T)\left(\theta_1^2+\delta_1^2\right)$ in the first sample, and $W(1-T)\left(\theta_2^2+\delta_2^2\right)$ in the second sample. This shows that, in both the subsamples, Gupta et al. [13] ORRM is more protective compared to Mehta et al. [14] ORRM. The measures of privacy for Huang [12] ORRM are given by $\left(\mu_X^2+\sigma_X^2\right)W\gamma_1^2+W\left(\theta_1^2+\delta_1^2\right)$ and $\left(\mu_X^2+\sigma_X^2\right)W\gamma_2^2+W\left(\theta_2^2+\delta_2^2\right)$ in the first and second subsamples, respectively. The measures of privacy for the proposed estimator in split sample approach are the same as that of Mehta et al. [14] ORRM. In double response approach the measure of privacy is given by $\frac{W(1-T-F)}{4}\left(\theta_1^2+\theta_2^2+\delta_1^2+\delta_2^2-2\theta_1\theta_2\right)$ which is equal to measure of privacy provided by Mehta et al. [14] ORRM if and only if $3\left(\theta_1^2+\delta_1^2\right)=\left(\theta_2^2+\delta_2^2-2\theta_1\theta_2\right)$ or $3E\left(D_1^2\right)=\left\{E\left(D_2^2\right)-2E(D_1)E(D_2)\right\}$. This shows that the proposed double response approach may be made more protective compared to Mehta et al. [14] ORRM at the cost of increased variance. In fact, it is a trade-off between the efficiency and privacy protection. That is, we can have highly efficient estimator by compromising on privacy. Similarly, we can build a more protective model by compromising on the efficiency.

Efficiency Comparison

We compare the proposed split sample and double response approaches with the Mehta et al. [14], Huang [12] and Gupta et al. [13] ORRMs in terms of relative efficiency.

(i) $\hat{\mu}_{XZ}$ versus $\hat{\mu}_{XM}$ and $\hat{W}_Z$ versus $\hat{W}_M$

The proposed estimators $\hat{\mu}_{XZ}$ and $\hat{W}_Z$ are relatively more efficient than the corresponding estimators $\hat{\mu}_{XM}$ and $\hat{W}_M$ of Mehta et al. [1] if $Var(\hat{\mu}_{XM})\geq Var(\hat{\mu}_{XZ})$ and $Var\left(\hat{W}_M\right)\geq Var\left(\hat{W}_Z\right)$. Since $\sigma^2_{R_i}=\sigma^2_{Z_i}$, from (4), (5), (20) and (21), it is easy to show that $\hat{\mu}_{XZ}$ and $\hat{W}_Z$ are relatively more efficient than $\hat{\mu}_{XM}$ and $\hat{W}_M$ if

$$\frac{(\theta_2+\theta_1)^2}{(\theta_2-\theta_1)^2}>1,$$

which is always true for every value of θ_1 and θ_2.

(ii) $\hat{\mu}_{XZ}$ versus $\hat{\mu}_{XG}$ and $\hat{\mu}_{XH}$

The proposed estimator $\hat{\mu}_{XZ}$ is relatively more efficient than $\hat{\mu}_{XG}$ and $\hat{\mu}_{XH}$ if $Var(\hat{\mu}_{XG})\geq Var(\hat{\mu}_{XZ})$ and $Var(\hat{\mu}_{XH})\geq Var(\hat{\mu}_{XZ})$. From (9), (14) and (21), we see that it is difficult to derive the efficiency conditions for $\hat{\mu}_{XZ}$. We calculated the relative efficiency numerically through simulations by defining $RE_1=\frac{Var(\hat{\mu}_{XG})}{Var(\hat{\mu}_{XZ})}$ and $RE_2=\frac{Var(\hat{\mu}_{XH})}{Var(\hat{\mu}_{XZ})}$. For a simulation study, we fixed

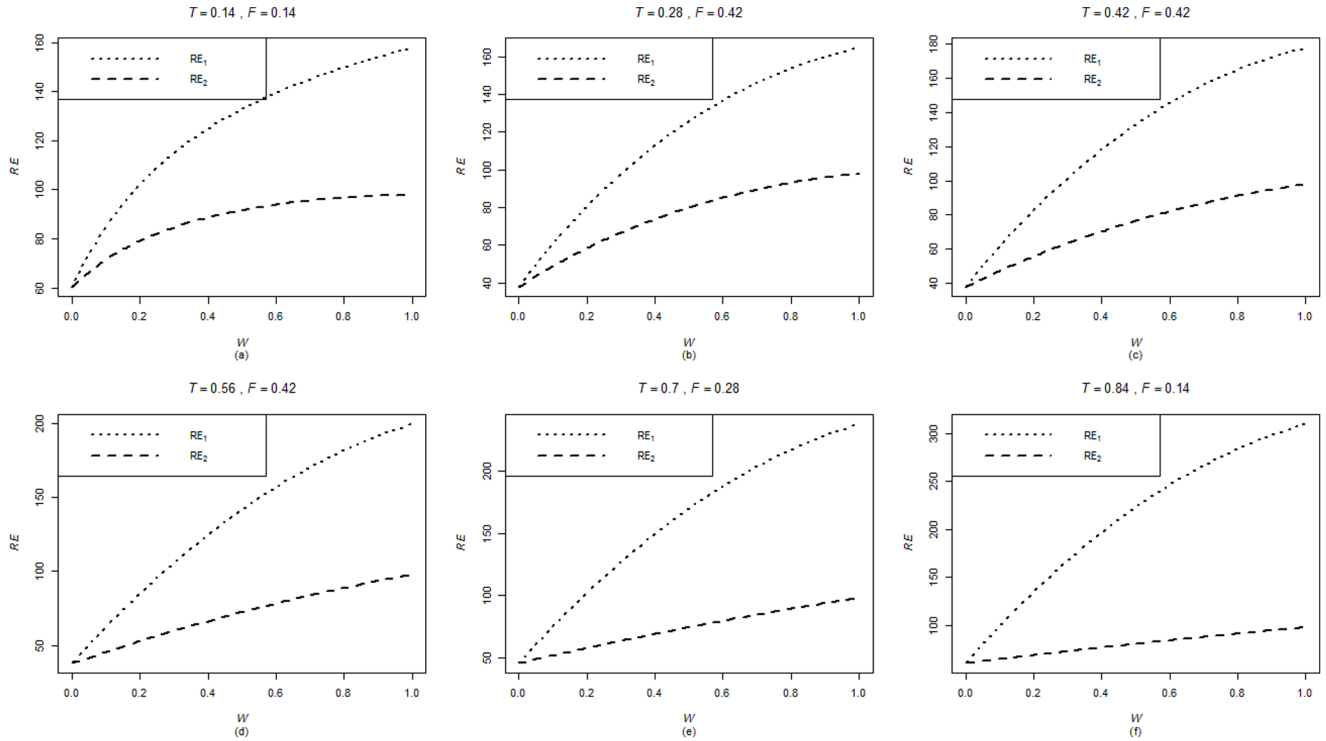

Figure 1. RE_1 **and** RE_2 **for** $(\mu_X^2,\sigma_X^2)=(1,1)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** $(\gamma_1^2,\gamma_2^2)=(2,3)$.
doi:10.1371/journal.pone.0083557.g001

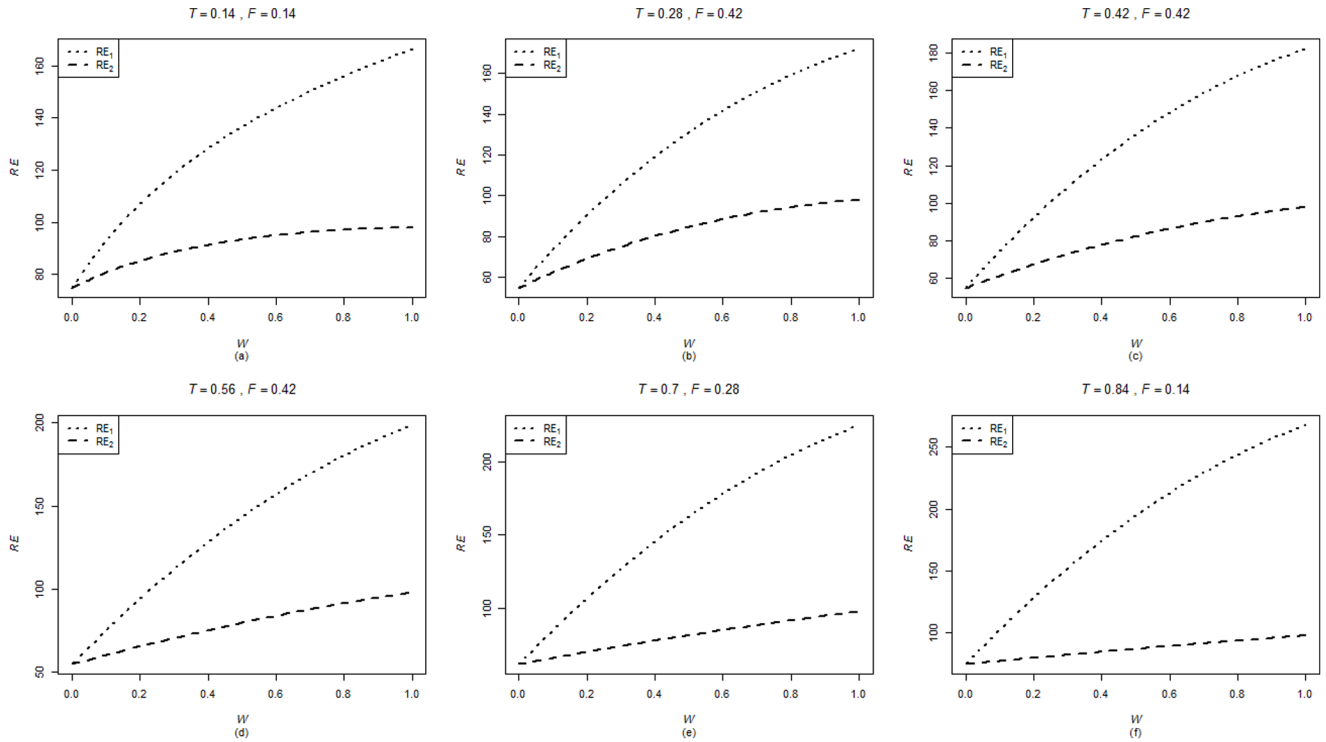

Figure 2. RE_1 **and** RE_2 **for** $(\mu_X^2,\sigma_X^2)=(1,2)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** $(\gamma_1^2,\gamma_2^2)=(2,3)$.
doi:10.1371/journal.pone.0083557.g002

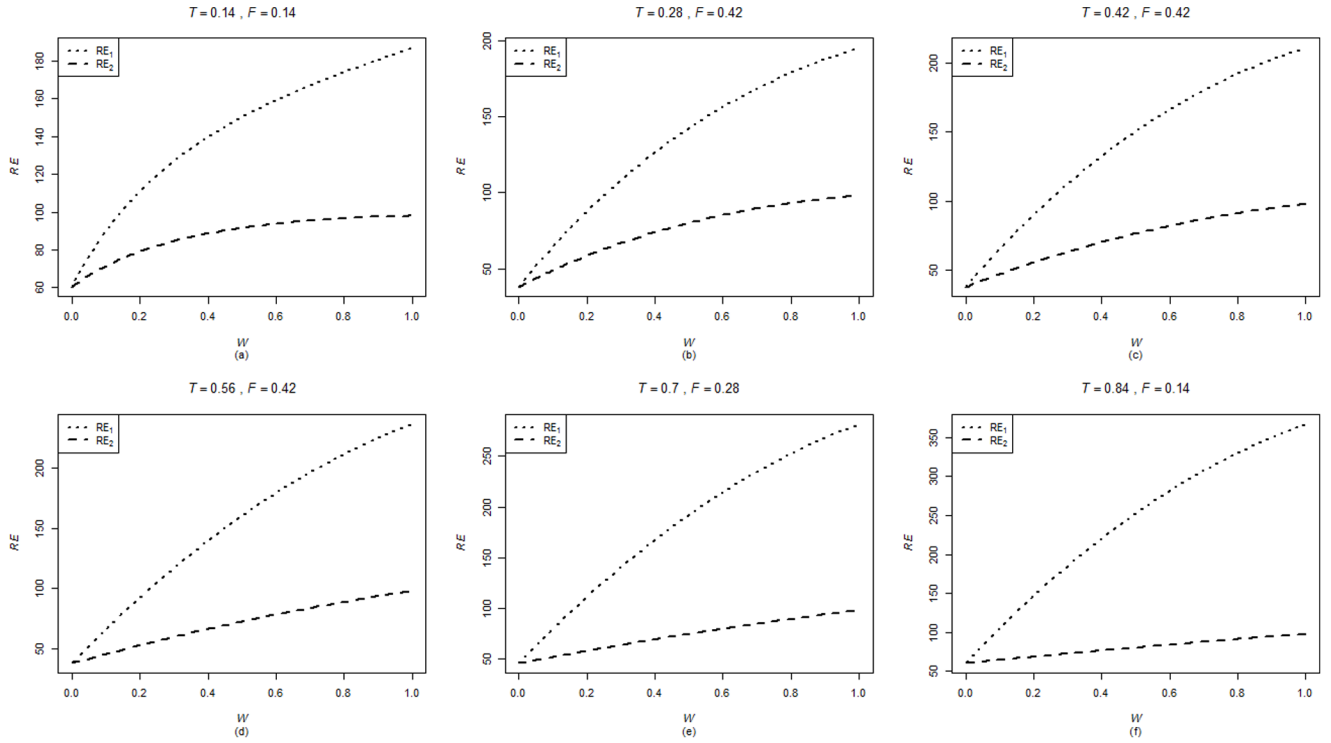

Figure 3. RE_1 **and** RE_2 **for** $(\mu_X^2,\sigma_X^2)=(1,2)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** $(\gamma_1^2,\gamma_2^2)=(2,3)$.
doi:10.1371/journal.pone.0083557.g003

$n_1=n_2=25$. We assumed that $X\sim N(\mu_X,\sigma_X^2)$, $D_i\sim N(\theta_i,\delta_i^2)$ and $S_i\sim N(1,\gamma_i^2)$, $i=1,2$. To simulate the data from the first subsample, we generated $n_1=25$values from a Bernoulli variable, say Q, with the parameter $\{F+(1-T-F)W\}$, where F, T and W are known. We, then, generated $n_1=25$ random values each

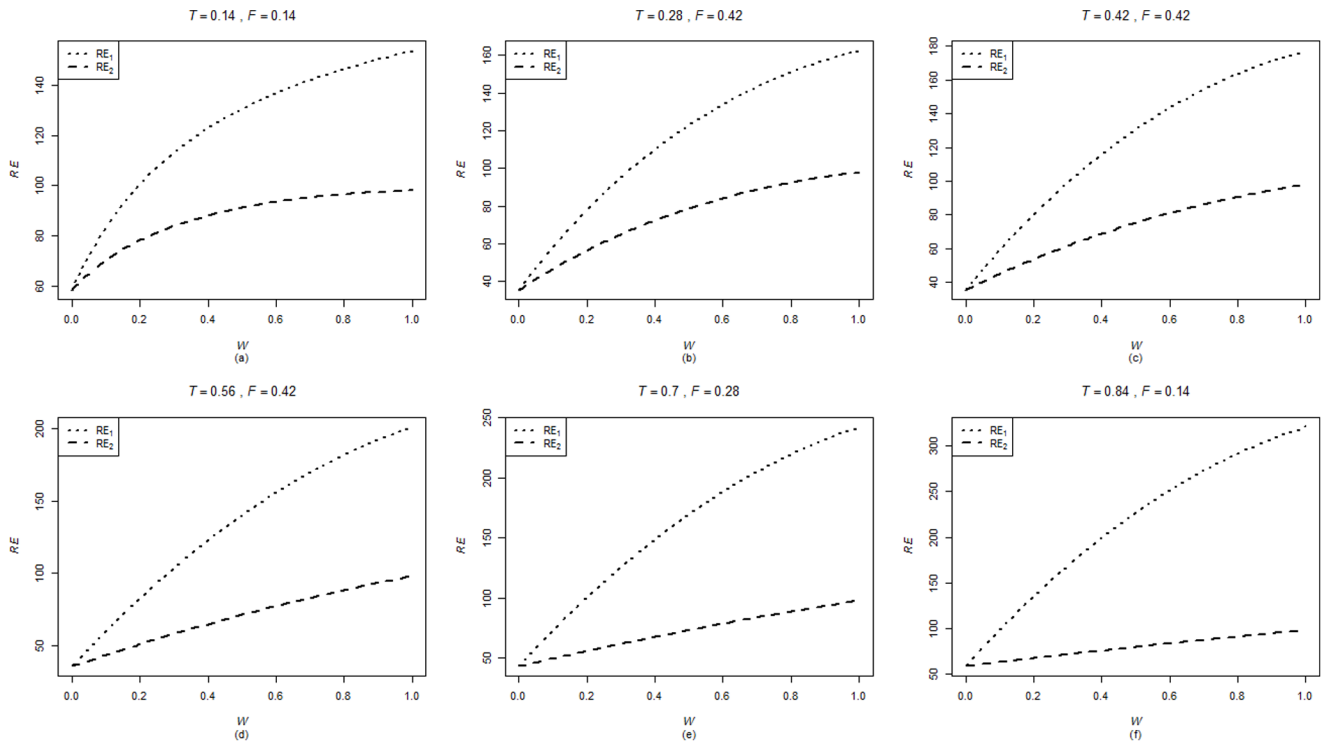

Figure 4. RE_1 **and** RE_2 **for** $(\mu_X^2,\sigma_X^2)=(1,1)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** ,$(\gamma_1^2,\gamma_2^2)=(2,4)$.
doi:10.1371/journal.pone.0083557.g004

on the variables X and D_1 from $X \sim N(\mu_X, \sigma_X^2)$ and $D_i \sim N(\theta_i, \delta_i^2)$, respectively. We took $R_{1j} = X_j$ if $Q = 0$, and $R_{1j} = X_j + D_{1j}$, otherwise. Similarly, $n_2 = 25$ values of R_2 from the second subsample are generated as: $R_{2j} = X_j$ if $Q = 0$, and $R_{2j} = X_j - D_{2j}$, otherwise. Same algorithm is used to generate the values of Z'_{ij} and $Z''_{ij}, (i = 1,2, j = 1,2,...25)$. Once the data have been generated, different estimators $(\hat{\mu}_{XZ}, \hat{\mu}_{XG}, \hat{\mu}_{XH})$ are computed using the corresponding formulae in (7), (12) and (20). The variances of these estimators are obtained using 5000 iterations. The relative efficiency results (for the different scenarios given below) are given in the Figures 1–4.

(iii) $\hat{\mu}'_{XZ}$ versus $\hat{\mu}_{XG}$ and $\hat{\mu}_{XH}$

The proposed estimator $\hat{\mu}'_{XZ}$ is relatively more efficient than $\hat{\mu}_{XG}$ and $\hat{\mu}_{XH}$ if $Var(\hat{\mu}_{XG}) \geq Var(\hat{\mu}'_{XZ})$ and $Var(\hat{\mu}_{XH}) \geq Var(\hat{\mu}'_{XZ})$. From (9), (14) and (30), we see that it is difficult to derive the efficiency conditions for $\hat{\mu}_{XZ}$. We, again, calculated the relative efficiency numerically through simulations by defining $RE_3 = \frac{Var(\hat{\mu}_{XG})}{Var(\hat{\mu}'_{XZ})}$ and $RE_4 = \frac{Var(\hat{\mu}_{XH})}{Var(\hat{\mu}'_{XZ})}$. We used the similar algorithm to simulate the values of R'_{ij}, Z'_{ij} and Z''_{ij}. It is to be noted that we simulated $n_1 + n_2 = 50$ values of $R'_i (i = 1,2.)$ and 25 values each of Z'_i and $Z''_i (i = 1,2.)$. The relative efficiency results are given in the Figures 5–8.

To calculate RE_1, $RE_2 RE_3$ and RE_4,we take the following different scenarios:

a. $(\mu_X^2, \sigma_X^2) = (1,1)$, $(\theta_1^2, \theta_2^2) = (2,3)$, $(\delta_1^2, \delta_2^2) = (1,1)$, $(\gamma_1^2, \gamma_2^2) = (2,3)$

b. $(\mu_X^2, \sigma_X^2) = (1,2)$, $(\theta_1^2, \theta_2^2) = (2,3)$, $(\delta_1^2, \delta_2^2) = (1,1)$, $(\gamma_1^2, \gamma_2^2) = (2,3)$

c. $(\mu_X^2, \sigma_X^2) = (2,1)$, $(\theta_1^2, \theta_2^2) = (2,3)$, $(\delta_1^2, \delta_2^2) = (1,1)$, $(\gamma_1^2, \gamma_2^2) = (2,3)$

d. $(\mu_X^2, \sigma_X^2) = (1,1)$, $(\theta_1^2, \theta_2^2) = (2,3)$, $(\delta_1^2, \delta_2^2) = (1,1)$, $(\gamma_1^2, \gamma_2^2) = (2,4)$,

and study the effect of γ_1^2 and γ_2^2 on RE_1, RE_2, RE_3 and RE_4. The relative efficiencies are calculated for different values of T and F over the whole range of W. It is observed that the proposed estimator $\hat{\mu}_{XZ}$ performs better (in terms of relative efficiency) than the $\hat{\mu}_{XG}$ and $\hat{\mu}_{XM}$. Also, the proposed estimator $\hat{\mu}'_{XZ}$ performs relatively better than $\hat{\mu}_{XG}$ and $\hat{\mu}_{XH}$. It can easily be verified through simulations that RE_1, RE_2, RE_3 and RE_4 are independent of $n_1 = n_2$. To save the space we have not presented the graphs for varying values of $n_1 = n_2$. From Figures 1–8 following observations are made.

(i) RE_1, RE_2, RE_3 and RE_4 are not seriously affected by the difference between γ_1^2 and γ_2^2 when the other parameters are fixed (see Figures 1 and 4 or 5 and 8).

(ii) RE_1, RE_2, RE_3 and RE_4 increase, over the whole range of W, with an increase in T when the other parameters, except F, are kept fixed (see Figures 1–8).

(iii) RE_1, RE_2, RE_3 and RE_4 are not seriously affected by change in μ_X^2 and/or σ_X^2 (see Figures 1 and 3, and 5 and 7 or 1 and 2 and 5 and 6).

(iv) Split sample approach is more efficient than double response approach

(v) The proposed estimators of mean through split sample and double response approaches do not need a smaller values of T irrespective of the sensitivity level W and the forced scrambling parameter F.

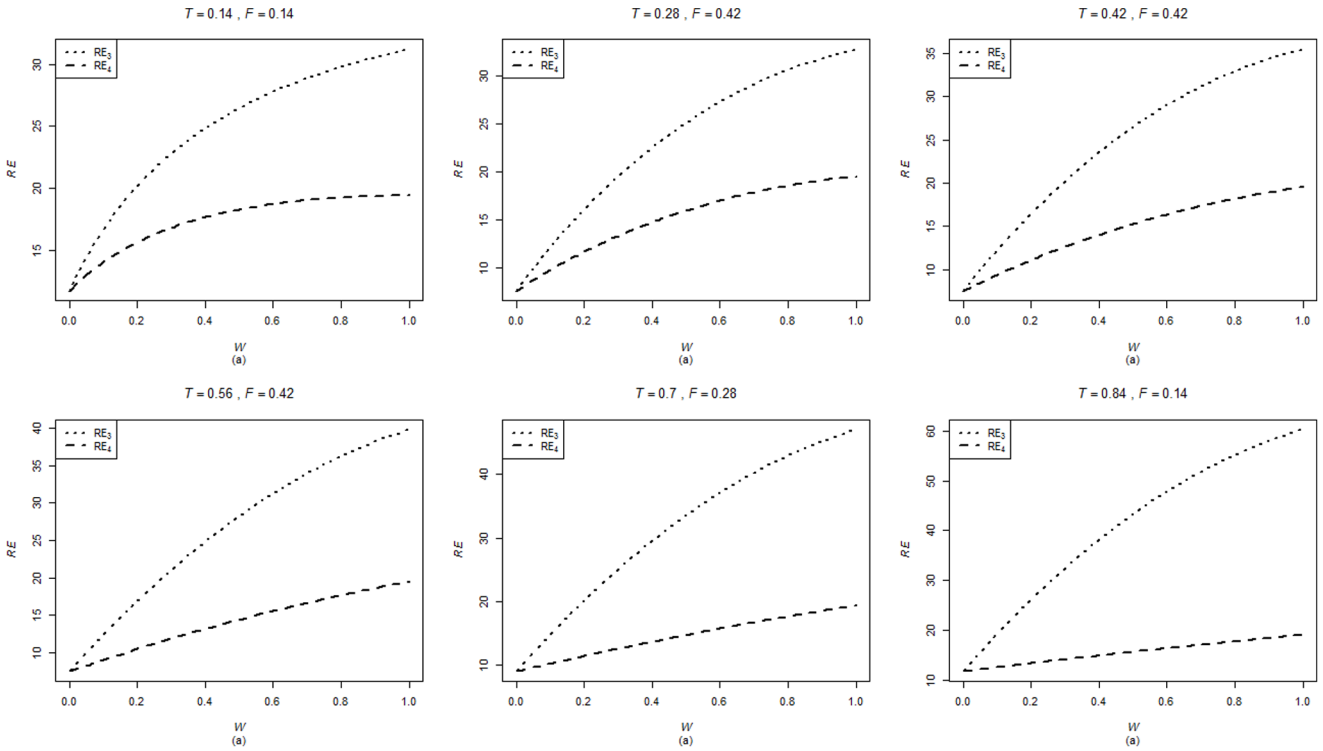

Figure 5. RE_3 **and** RE_4 **for** $(\mu_X^2, \sigma_X^2) = (1,1)$, $(\theta_1^2, \theta_2^2) = (2,3)$, $(\delta_1^2, \delta_2^2) = (1,1)$ **and,** $(\gamma_1^2, \gamma_2^2) = (2,3)$.
doi:10.1371/journal.pone.0083557.g005

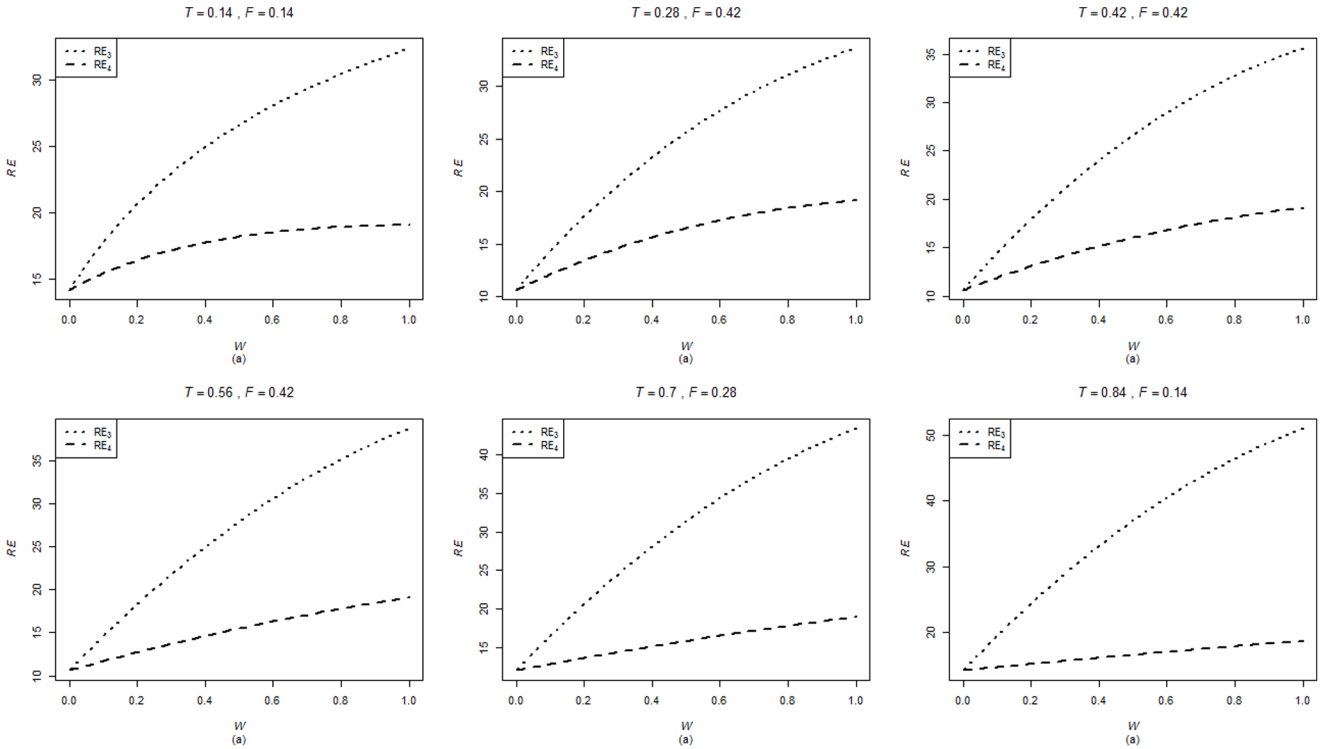

Figure 6. RE_3 **and** RE_4 **for** $(\mu_X^2,\sigma_X^2)=(1,2)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** $(\gamma_1^2,\gamma_2^2)=(2,3)$.
doi:10.1371/journal.pone.0083557.g006

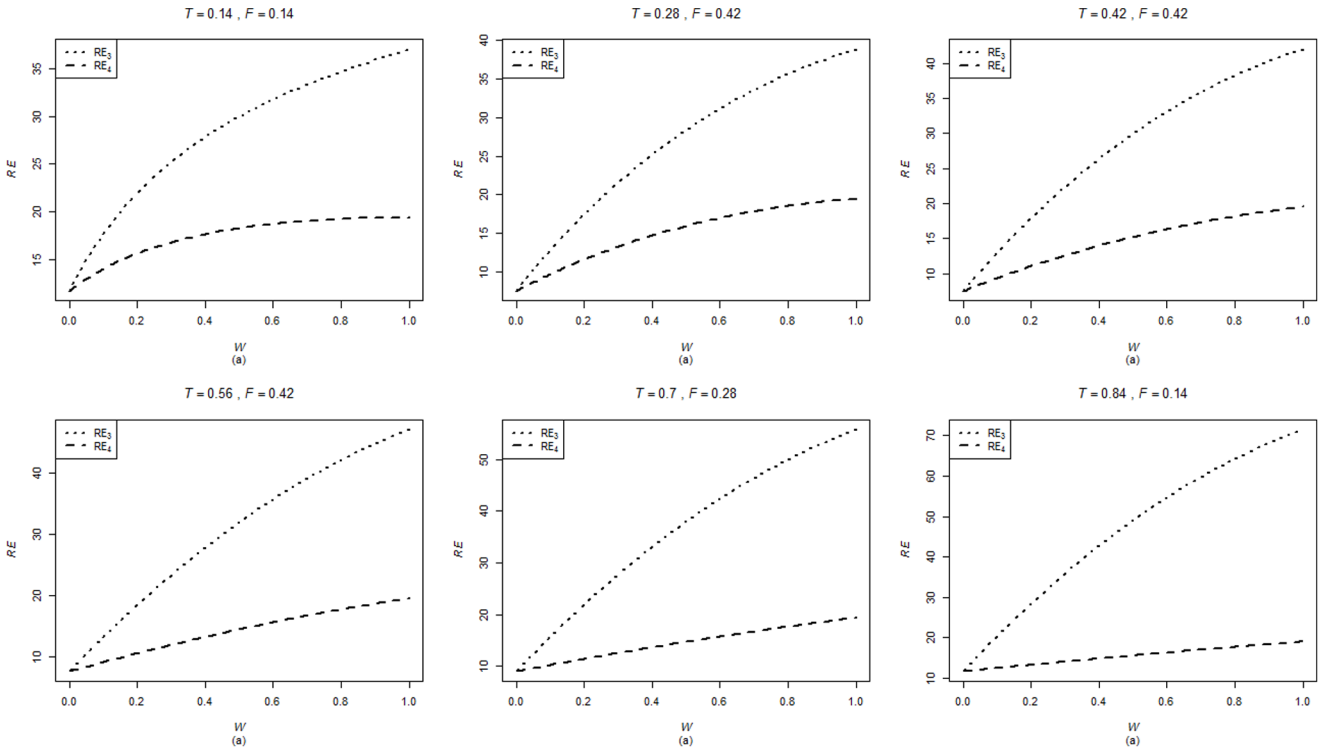

Figure 7. RE_3 **and** RE_4 **for** $(\mu_X^2,\sigma_X^2)=(2,1)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** $(\gamma_1^2,\gamma_2^2)=(2,3)$.
doi:10.1371/journal.pone.0083557.g007

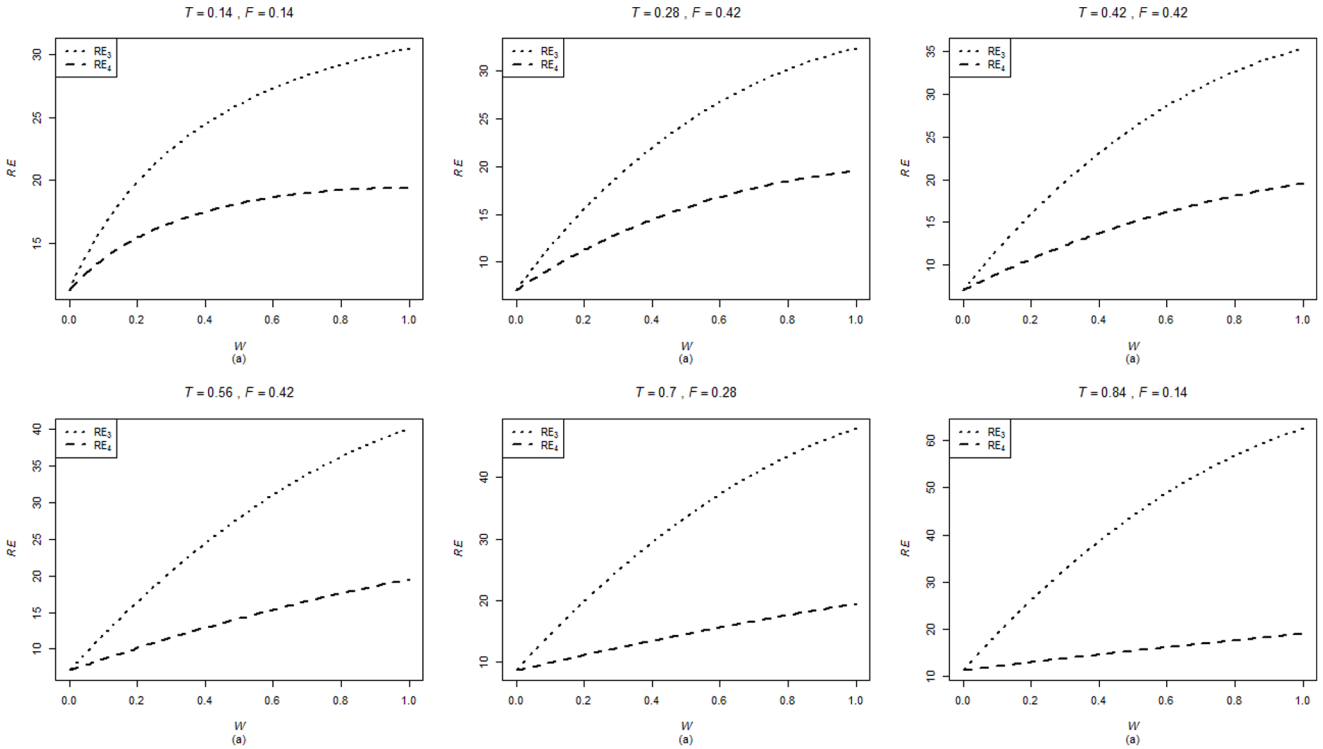

Figure 8. RE_3 **and** RE_4 **for** $(\mu_X^2,\sigma_X^2)=(1,1)$, $(\theta_1^2,\theta_2^2)=(2,3)$, $(\delta_1^2,\delta_2^2)=(1,1)$ **and** $(\gamma_1^2,\gamma_2^2)=(2,4)$.
doi:10.1371/journal.pone.0083557.g008

Conclusion

To estimate the mean, variance and the sensitivity level of a sensitive variable optional randomized response model by Mehta et al. [14] is improved. Utilizing the idea of additive scrambling in one sample and subtractive scrambling in the other subsample, we have proposed unbiased estimators of mean, variance and sensitivity level. We compared the proposed procedure with Mehta et al. [14] Huang [12], and Gupta et al. [13] procedure. The proposed idea resulted in the improved estimation of mean of the study variable. It has been shown by Huang [12] that his procedure works better than Gupta et al. [4] procedure. Therefore, the proposed split sample procedure is also better than Gupta et al. [4] procedure both in terms of relative efficiency and providing unbiased estimators of the mean μ_X, sensitivity level W and variance σ_X^2 of the study variable. Like Huang [12], the proposed procedure has the same advantage of estimating the variance of Y with no bias. Unlike Gupta et al. [4], proposed procedures do not require larger value of truth parameter (T) when the study variable is highly sensitive. This may be considered the major advantage of the proposed procedures. It has been established that the proposed procedure of estimating mean is more efficient than all the procedures considered in this study. Moreover, as far as, the estimation of sensitivity is concerned we observed that the proposed estimators are less efficient (not shown in the figures) than all the estimators considered here except Mehta et al. [14].

As a final comment, we recommend using proposed procedures in the field surveys without increasing sampling cost when estimation of mean of the study variable is of prime interest.

Author Contributions

Conceived and designed the experiments: ZH BA MA. Performed the experiments: ZH BA MA. Analyzed the data: ZH BA MA. Wrote the paper: ZH.

References

1. Warner SL (1965) Randomized response: a survey for eliminating evasive answer bias. Journal of the American Statistical Association 60: 63–69.
2. Greenberg BG, Kubler RR, Horvitz DG (1971) Applications of RR technique in obtaining quantitative data. Journal of the American Statistical Association 66: 243–250.
3. Eichhorn BH, Hayre LS (1983) Scrambled randomized response methods for obtaining sensitive question data. Journal of Statistical Planning and Inference 7: 307–316.
4. Gupta S, Gupta B, Singh S (2002) Estimation of sensitivity level of personal interview survey questions. Journal of Statistical Planning and Inference 100: 239–247.
5. Gupta S, Shabbir J (2004) Sensitivity estimation for personal interview survey Questions. Statistica, anno LXIV, n, 4: 643–653.
6. Bar-Lev SK, Bobovitch E, Boukai B (2004) A note on randomized response models for quantitative data. Metrika 60: 255–260.
7. Gupta SN, Thornton B, Shabbir J, Singhal S (2006) A Comparison of Multiplicative and Additive Optional RRT Models, Journal of Statistical Theory and Applications 5: 226–239.
8. Hussain Z, Shabbir J (2007) Estimation of mean of a sensitive quantitative variable. Journal of Statistical research 41(2), 83–92.
9. Saha A (2008) A randomized response technique for quantitative data under unequal probability sampling. Journal of Statistical Theory and Practice 2(4): 589–596.
10. Chaudhuri A (2012) Unbiased estimation of sensitive proportion in general sampling by three non randomized response techniques. Journal of Statistical Theory and Practice 6(2), 376–381.
11. Hussain Z, Shabbir J (2013) Estimation of the mean of a socially undesirable characteristic. Scientia Iranica E (20)3: 839–845.
12. Huang KC (2010) Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling. Metrika 71: 341–352.

13. Gupta S, Shabbir J, Sehra S (2010) Mean and sensitivity estimation in optional randomized response models. Journal of Statistical Planning and Inference 140(10), 2870–2874.
14. Mehta S, Dass BK, Shabbir J, Gupta S (2012) A three stage optional randomized response model. Journal of Statistical Theory and Practice 6(3), 417–426.
15. Gupta S, Shabir J, Sousa R, Corte-Real P (2012) Estimation of the mean of a sensitive variable in the presence of auxiliary information. Communications in Statistics-Theory and Methods 41: 2394–2404.
16. Gupta S, Mehta S, Shabbir J, Dass BK (2011) Some optimatility issues in estimating two-stage optional randomized response models. American Journal of Mathematical and Management Sciences 31(1–2), 1–12.
17. Murthy MN (1967) Sampling Theory and Methods. Calcutta, India. Statistical Publishing Society.
18. Zaizai Y, Jingyu W, Junfeng L (2009) An efficiency and protection degree based comparison among the quantitative randomized response strategies. Communications in Statistics- Theory and Methods 38 (3), 400–408.